Nucleic Acids Research, 2004, Vol. 32, Database issue ***D125–D128***
DOI: 10.1093/nar/gkh130

THGS: a web-based database of Transmembrane Helices in Genome Sequences

S. A. Fernando[1], P. Selvarani[1], Soma Das[1], Ch. Kiran Kumar[1], Sukanta Mondal[2], S. Ramakumar[1,2,3] and K. Sekar[1,3,*]

[1]Bioinformatics Centre, [2]Department of Physics, [3]Supercomputer Education and Research Centre, Indian Institute of Science, Bangalore 560 012, India

Received August 29, 2003; Revised and Accepted October 27, 2003

ABSTRACT

Transmembrane Helices in Genome Sequences (THGS) is an interactive web-based database, developed to search the transmembrane helices in the user-interested gene sequences available in the Genome Database (GDB). The proposed database has provision to search sequence motifs in transmembrane and globular proteins. In addition, the motif can be searched in the other sequence databases (Swiss-Prot and PIR) or in the macromolecular structure database, Protein Data Bank (PDB). Further, the 3D structure of the corresponding queried motif, if it is available in the solved protein structures deposited in the Protein Data Bank, can also be visualized using the widely used graphics package RASMOL. All the sequence databases used in the present work are updated frequently and hence the results produced are up to date. The database THGS is freely available via the world wide web and can be accessed at http://pranag.physics.iisc.ernet.in/thgs/ or http://144.16.71.10/thgs/.

INTRODUCTION

The prediction of transmembrane helices in integral membrane proteins is an important aspect of structural genomics. Several research groups working in the area of genome annotation, target-receptor isolation and characterization and target specific pharmacological drug design are interested in identifying membrane-associated proteins from primary structure. Computational methods have proved to be good and will continue to be one of the most efficient tools for the analysis of transmembrane proteins.

Transmembrane proteins are integral membrane proteins that interact exclusively with the hydrophobic tails of the lipid bilayer. These proteins either span the bilayer or are embedded near the hydrophobic polar head groups. The membrane-spanning α-helical domains are embedded in membranes by hydrophobic interactions with the lipid interior of the bilayer and by ionic interactions with the polar head groups of the phospholipids. α-helical composition makes up a significant percentage of the transmembrane domains in these proteins. Most transmembrane proteins are formed from bundles of helices that traverse the membrane lipid bilayer and are known to be the most commonly occurring secondary structural elements in membrane proteins (1,2). Many groups of membrane proteins, including ion channels, toxins, antibiotics and receptors have α-helical structure (3). It was observed that over 30% of membrane proteins in known genomes are twisted into α-helices.

In general, the transmembrane helical regions comprise a region of 18 or more amino acids and most of them are hydrophobic in nature. The most abundant amino acids in transmembrane regions are leucine, isoleucine, valine, phenylalanine, alanine, glycine, serine and threonine (4). These amino acids together constitute almost 75% in transmembrane regions (5). The amino acid pattern in the transmembrane region is usually GXXXG (where G is glycine and X is any amino acid), and is an important criterion in helix–helix interactions (6,7).

The available gene sequences are increasing rapidly with time and in order to analyse the large volume of database, it is always essential to have an automated search engine. To the best of our knowledge, there is no web-based database to identify transmembrane and globular proteins in a given organism available in the genome database. Towards this effort, several packages have been developed for identifying transmembrane proteins in genome sequences. By using a transmembrane topology prediction method, TMHMM (8), genome-scale analysis of proteins has been carried out in this database. TMHMM is a program that predicts and characterizes transmembrane domains using Hidden Markov Models. This program depicts different protein sequence structures, including the helix core, inside and outside loops (short and long), helix caps (C and N) and globular domains. TMHMM can also locate and orient the domains (topology) with the membranes that they transverse. The program TMHMM used in THGS to predict transmembrane helices is more accurate (77–80%) compared with other programs available in the literature.

ABOUT THGS

THGS is an interactive web-based database to delineate between the transmembrane proteins and globular proteins of

*To whom correspondence should be addressed. Tel: +91 80 3601409; Fax: +91 80 3600683; Email: sekar@physics.iisc.ernet.in

the organisms whose complete genome sequences are available in the Genome Database. It can also distinguish between signal peptides and membrane proteins using the prediction method SignalP (9). Initially, the transmembrane proteins and globular proteins are computed using the program TMHMM and are included in the THGS database under different clusters. The proposed database (THGS) is developed in such a way that it can display the transmembrane and globular proteins for all or for a particular organism. In the case of transmembrane proteins, THGS uses different colour schemes to display the residues in the helices (red), cytoplasm (green) and outside the cell (yellow). In addition, the percentages of amino acids in different regions are shown in a graphical format.

UTILITIES OF THGS

The data are organized in a format appropriate for a quick search in the database. To speed up the computation, derived databases have been created and deployed for options like isoelectric point, sequence length and molecular weight. The THGS database provides several options for the user to search and analyse gene sequences of interest.

The following are the search options available for users in THGS: (i) display transmembrane proteins; (ii) display globular proteins; (iii) information about complete genomes; (iv) search using number of helices; (v) search using number of residues in a helical region; (vi) search using sequence length; (vii) search using isoelectric point; (viii) search using molecular weight; (ix) search using protein name; (x) identical pattern matching; (xi) similar pattern matching.

Different colour codes have been adopted to delineate different regions (transmembrane helices, cytoplasmic and extracellular domains). THGS has an additional feature for motif search in sequence and structural databases, when a query motif is submitted for search across PIR, Swiss-Prot and PDB. Further, the 3D structure of the input or the query motif, if available in the PDB, can be viewed using the graphic display package RASMOL (10). The user can also extract the atomic coordinates of the motif or part of the protein for further analysis. RASMOL installation instructions are provided (see http://144.16.71.10/thgs/rasmol.html for instructions). The search engine THGS is available on the world wide web and users can easily submit their queries. In trial runs, the output page appeared in ~10–15 s depending on the nature of the query and network traffic.

DATA ORGANIZATION

THGS uses six different databases, namely, (i) Genome Database, (ii) Swiss-Prot (11), (iii) PIR (12), (iv) PDB (13), (v) 25% and (vi) 90% non-homologous protein structures (14). The Genome database used in the present software has been downloaded from the National Centre for Biotechnology Information (NCBI). The PDB is maintained by the Research Collaboratory for Structural Bioinformatics (RCSB), University of California, San Diego, CA, USA. These two (GDB and PDB) databases are also available in the form of an anonymous FTP server at the Bioinformatics Centre, Indian Institute of Science, Bangalore, India. The sequence database Swiss-Prot contains 135 493 entries (as on October 2003). The PIR contains 283 347 non-redundant entries (as on 25 August 2003). The non-homologous data set (both 25% and 90%) was derived by Hobohm and Sander. The complete genome sequences of the following organisms, namely, (i) *Anopheles gambiae* (ii) *Arabidopsis thaliana*, (iii) bacteria, (iv) *Caenorhabditis elegans*, (v) *Drosophila melanogaster*, (vi) *Plasmodium falciparum*, (vii) *Saccharomyces cerevisiae* and (viii) *Schizosaccharomyces pombe* are available via the NCBI genome database anonymous FTP site. The search engine is written using CGI/PERL scripts. The front-end input data part of this tool is coded in HTML and JavaScript and allows user-friendly web forms. The software is easy to use and tested on Windows 95/98/2000, Windows NT, Linux and Silicon Graphics (SGI) platforms through the most popular web browsers Netscape and Internet Explorer. The above-mentioned facility is freely available over the world wide web (www) at the URL http://pranag.physics.iisc.ernet.in/thgs/ or http://144.16.71.10/thgs/.

CASE STUDY

Figure 1 shows the output of the result of a typical search for the *Mycobacterium tuberculosis* H37RV genome sequence. It shows, for a particular hypothetical protein, the transmembrane proteins with the helical region sequences in red, cytoplasmic amino acid sequences in green and extracellular residues in yellow. In addition, a detailed analysis of the amino acid residues will be displayed in a separate window on clicking the button 'More information'. This option shows a table containing information about the total number of occurrences and percentage of individual amino acids along with a graphical representation. An additional button 'Signal Peptide' predicts the presence of signal peptides, if any, and the location of cleavage sites in a specific protein sequence.

The option 'Analysis of transmembrane proteins in the available GDB' (Fig. 2) displays the complete information about a particular organism of interest. Figure 2 shows the information about *M.tuberclulosis* H37Rv. The graphics panel shown in Figure 2 can be displayed on clicking the button 'Graph' provided at the end of the output frame.

A sample of a typical search for a conserved structural motif (KMSKS) in the superfamily of nucleotide binding proteins in all the genome sequences available in the Genome Database (figure not shown) is outlined here. This particular motif has 22 matches in transmembrane proteins and 803 matches in globular proteins. An additional option has been provided for users to see the occurrence of the given sequence motif in the other sequence (Swiss-Prot and PIR) and structural (PDB) databases. The same sequence motif 'KMSKS' occurs 419 times in Swiss-Prot, 386 times in PIR and 39 times in PDB.

The above outlined database provides several useful utilities to users working in the area of structural genomics and functional proteomics. The databases used in THGS will be updated from their respective primary servers at regular intervals and hence the results produced by THGS are up to date. New features will be added based on requests from the scientific community around the world. Users of THGS are requested to cite this article when referencing the database. Questions, comments and suggestions can be sent to Dr K. Sekar at sekar@physics.iisc.ernet.in.

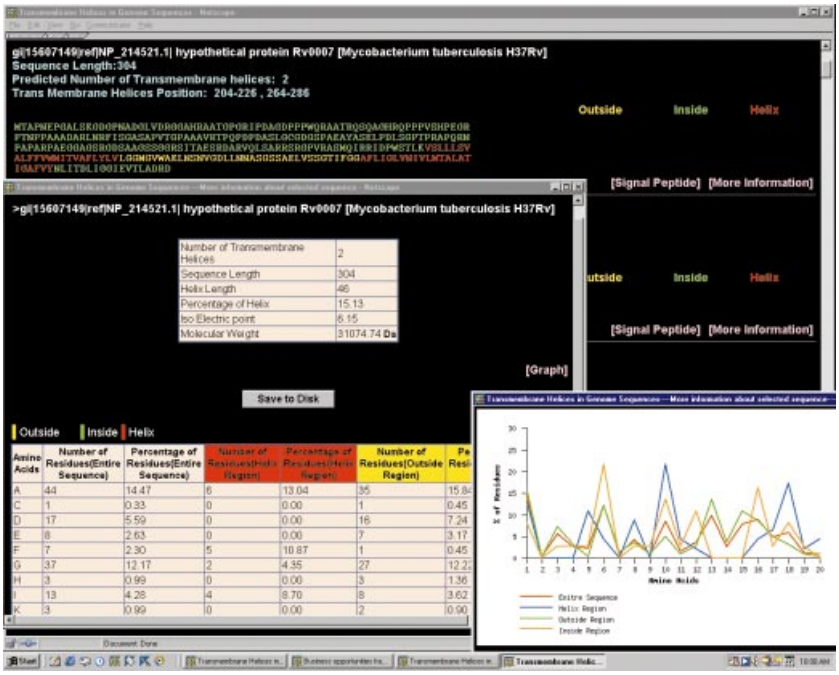

Figure 1. This page depicts the output of the program THGS for the gene gi|15607149|ref|NP_214521.1 (hypothetical protein Rv0007) of the organism *M.tuberculosis* H37Rv which is invoked using 'To display the transmembrane proteins' option from the left panel. Colouring schemes have been used to delineate various regions.

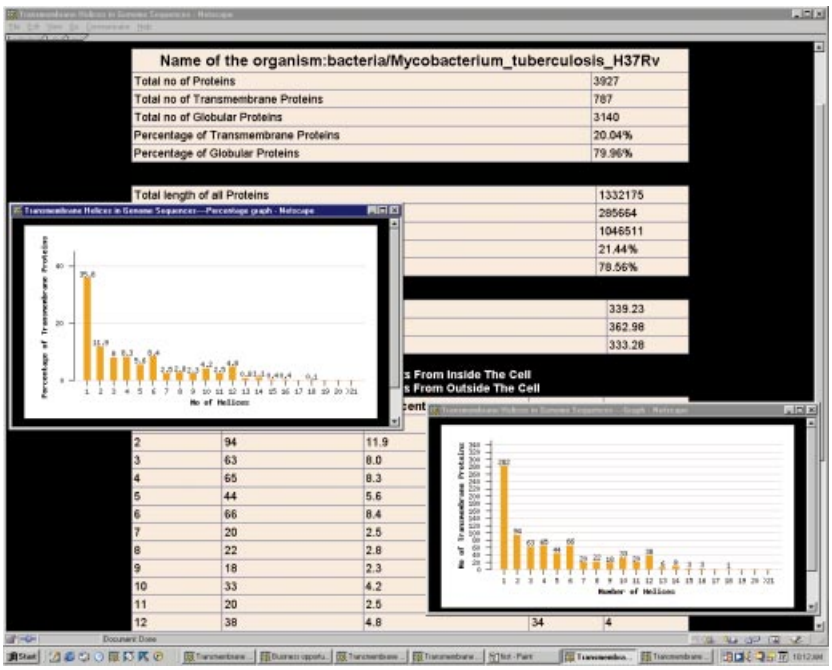

Figure 2. This is a typical output of the option 'Analysis of transmembrane proteins in the available genome sequences' for *M.tuberculosis* H37Rv. The graphics panel can be invoked by clicking the button 'graph' provided at the end of the output frame. [The graphs shown in Figure 1 and above are produced using the Ploticus data display program (15).]

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the use of the Bioinformatics Centre (DIC), Interactive Graphics-based Molecular Modelling (IGBMM) facility and the Supercomputer Education and Research Centre (SERC). The DIC and IGBMM facilities are supported by the Department of Biotechnology (DBT), Government of India, India. This work is supported by the Institute-wide computational Genomics program, supported by the DBT.

REFERENCES

1. Cohen,C. and Parry,D.A.D. (1990) α-Helical coiled coils and bundles: how to design an α-helical protein. *Proteins*, **7**, 1–15.
2. Cohen,C. and Parry,D.A.D. (1994) α-Helical coiled coils: more facts and better predictions. *Science*, **263**, 488–489.
3. Dieckmann,G.R. and DeGrado,W.F. (1997) Modeling transmembrane helical oligomers. *Curr. Opin. Struct. Biol.*, **7**, 486–494.
4. Jones,D.T., Taylor,W.R. and Thornton,J.M. (1994) A mutation data matrix for transmembrane proteins. *FEBS Lett.*, **339**, 269–375.
5. Arkin,I.T. and Brunger,A.T. (1998) Statistical analysis of predicted transmembrane α-helices. *Biochim. Biophys. Acta*, **1429**, 113–128.
6. Senes,A., Gerstein,M. and Engelman,D.M. (2000) Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with β-branched residues at neighboring positions. *J. Mol. Biol.*, **296**, 921–936.
7. Russ,W.P. and Engelman,D.M. (2000) The GxxxG motif: a framework for transmembrane helix–helix association. *J. Mol. Biol.*, **296**, 911–919.
8. Krogh,A., Larsson,B., von Heijne,G. and Sonnhammer,E.L.L. (2001) Predicting transmembrane protein topology with a Hidden Markov Model: Application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.
9. Nielson,H., Engelbrecht,J., Brunak,S. and von Heijne,G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.*, **10**, 1–6.
10. Sayle,R.A. and Milner-Whilte,E.J. (1995) RASMOL: Biomolecular graphics for all. *Trends Biochem. Sci.*, **20**, 374–382.
11. Bairoch,A. and Apweiler,R. (1998) The SWISS-PROT protein sequence databank and its supplement TrEMBL in 1998. *Nucleic Acids Res.*, **26**, 38–42.
12. Barker,W.C., Garavelli,J.S., Haft,D.H., Hunt,L.T., Marzec,C.R., Orcutt,B.C., Srinivasarao,G.Y., Yeh,L.S.L., Ledley,R.S., Mewes,H.W. *et al.* (1998) The PIR—International protein sequence database. *Nucleic Acids Res.*, **28**, 27–32.
13. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
14. Hobohm,U. and Sander,C. (1994) Enlarged representative set of protein structures. *Protein Sci.*, **3**, 522–524.
15. Ploticus 2.04 data display engine. Software, documentation and examples. Copyright 1998–2002. Stephen C. Grubb.